Attribution of Predictive Uncertainties in Classification Models

Abstract

Predictive uncertainties in classification tasks are often a consequence of model inadequacy or insufficient training data. In popular applications, such as image processing, we are often required to scrutinise these uncertainties by meaningfully attributing them to input features. This helps to improve interpretability assessments. However, there exist few effective frameworks for this purpose. Vanilla forms of popular methods for the provision of saliency masks, such as SHAP or integrated gradients, adapt poorly to target measures of uncertainty. Thus, state-of-the-art tools instead proceed by creating *counterfactual* or *adversarial* feature vectors, and assign attributions by direct comparison to original images. In this paper, we present a novel framework that combines path integrals, counterfactual explanations and generative models, in order to procure attributions that contain few observable artefacts or noise. We evidence that this outperforms existing alternatives through quantitative evaluations with popular benchmarking methods and data sets of varying complexity.

1 INTRODUCTION

Model uncertainties often manifest aspects of a system or data generating process that are not exactly understood [Hüllermeier and Waegeman, 2021], such as the influence of model inadequacy or a lack of diverse and representative data used during training. The ability to quantify and attribute such uncertainties to their sources can help scrutinize aspects in the functioning of a predictive model, and facilitate interpretability or fairness assessments in important machine learning applications [Awasthi et al., 2021]. The process is especially relevant in Bayesian inferential settings, which find applications in domains such as natural language processing [Xiao and Wang, 2019], network analysis [Perez and Casale, 2021] or image processing [Kendall and Gal, 2017], to name only a few.

Thus, there exists a growing interest in methods for uncertainty estimation [e.g. Depeweg et al., 2018, Smith and Gal, 2018, Van Amersfoort et al., 2020, Tuna et al., 2021] for purposes such as procuring adversarial examples, active learning or *out-of-distribution* detection. Recent work has proposed mechanisms for the attribution of predictive uncertainties to input features, such as pixels in an image [Van Looveren and Klaise, 2019, Antoran et al., 2021, Schut et al., 2021], with the goal of complementing interpretability tools disproportionately centred on explaining model scores, and to improve transparency in deployments of predictive models. These methods proceed by identifying *counterfactual* (in-distribution) or *adversarial* (out-of-distribution) explanations, i.e. small variations in the value of input features which output new model scores with minimal uncertainty. This has helped understand the strengths and weaknesses of various models. However, the relative contribution of individual pixels to poor model performance is up to human guesswork, or assigned by plain comparisons between an image and its altered representation. We report that uncertainty attributions derived following these approaches perform poorly, when measured by popular quantitative evaluations of image saliency maps.

In this paper, our goal is to similarly map uncertainties in classification tasks to their origin in images, and to measure the relative contribution of each individual pixel. We show that popular attribution methods based on *segmentation* [Ribeiro et al., 2016], *resampling* [Lundberg and Lee, 2017a] or *path integrals* [Sundararajan et al., 2017] are easily re-purposed for this purpose. However, we evidence that naive applications of these approaches perform poorly. Thus, we present a new framework through a novel combination of path integrals, counterfactual explanations and generative models. Our approach is to attribute uncertainties by traversing a domain of integration defined in latent space, which connects a counterfactual explanation with its original im-

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

age. The integration is projected into the observable pixel space through a generative model, and starts at a reference point which bears no predictive uncertainty. Hence, *completeness* is satisfied and uncertainties are fully explained and decomposed over pixels in an image.

We note that relying on generative models has recently gained traction for interpretability and score attribution purposes [Lang et al., 2021]. Through our method, we show how to leverage these models in order to procure clustered saliency maps, which reduce the observable noise in vanilla approaches. Applied to uncertainty attribution tasks, the proposed approach outperforms vanilla adaptations of popular interpretability tools such as LIME [Ribeiro et al., 2016], SHAP [Lundberg and Lee, 2017a] or integrated gradients [Sundararajan et al., 2017], as well as *blur* and *guided* variants [Xu et al., 2020, Kapishnikov et al., 2021]. We further combine these methods with *Xrai* [Kapishnikov et al., 2019], a popular segmentation and attribution approach. The assessment[1] is carried out through both quantitative and qualitative evaluations, using popular benchmarking methods and data sets of varying complexity.

2 UNCERTAINTY ATTRIBUTIONS

Consider a classification task with a classifier $f : \mathbb{R}^n \times \mathcal{W} \to \Delta^{|\mathcal{C}|-1}$ of a fixed architecture. The weights $\boldsymbol{w} \in \mathcal{W}$ are presumed to be fitted to some available train data set $\mathcal{D} = \{\boldsymbol{x}_i, c_i\}_{i=1,2,\ldots}$. Thus, the function $f(\boldsymbol{x}) \equiv f(\boldsymbol{x}, \boldsymbol{w})$ maps feature vectors $\boldsymbol{x} \in \mathbb{R}^n$ to an element in the standard $(|\mathcal{C}| - 1)$-simplex, which represents membership probabilities across classes in a set $\mathcal{C}$. In the following, we are concerned with the *entropy* as a measure of predictive uncertainty, i.e.

$$H(\boldsymbol{x}) = -\sum_{c \in \mathcal{C}} f_c(\boldsymbol{x}) \cdot \log f_c(\boldsymbol{x}) \tag{1}$$

where $f_c(\boldsymbol{x})$ represents the predicted probability of class-c membership. In Bayesian settings, we often consider a posterior distribution $\pi(\boldsymbol{w}|\mathcal{D})$ over weights in the model, and the term (1) may further be decomposed into *aleatoric* and *epistemic* components [Kendall and Gal, 2017]. These represent different types of uncertainties, including inadequate data and inappropriate modelling choices. For simplicity in the presentation, we defer those details to Appendix A.

Popular *resampling* or *gradient*-based methods can easily be adapted in order to attribute measures of uncertainty such as $H(\boldsymbol{x})$ to input features in an image. This includes tools such as LIME [Ribeiro et al., 2016], SHAP [Lundberg and Lee, 2017a] or *integrated gradients* (IG) [Sundararajan et al., 2017]. In Figure 1, we show an example application of integrated gradients to *dogs versus cats* data (further examples

[1]Source code for reproducing our results can be found at the following link: *[removed for blind review]*

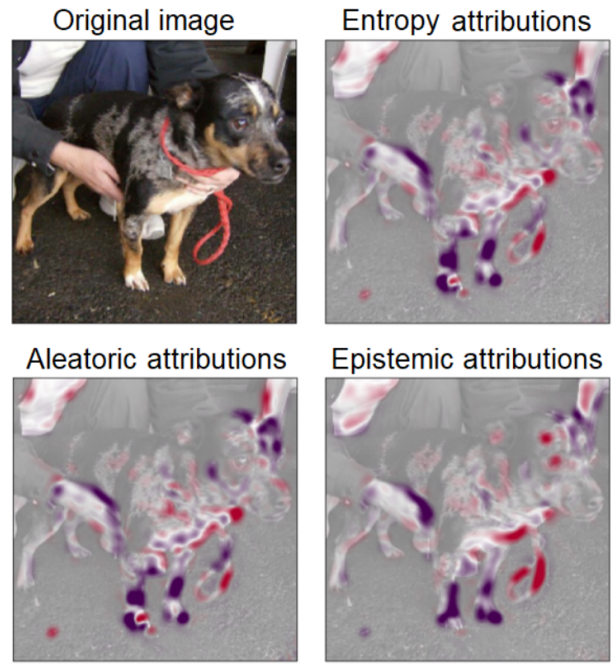

Figure 1: Example uncertainty attributions using integrated gradients. Classification task in *dogs versus cats* data. In red, positive attributions which *increase* entropy; in purple, negative attributions that *decrease* entropy.

are available in Appendix C). In the figure, the regions in red are identified as contributors to predictive uncertainties. We readily comprehend why the model struggles to predict any single class, by observing that a leash and a human hand are problematic. To the best of our knowledge, no research has yet explored the possibility of using these attribution methods to identify sources of uncertainty. Nevertheless, quantitative evaluations presented in Section 4 show that this approach offers generally poor performance.

2.1 PATH INTEGRALS

For later reference, we illustrate the above uncertainty attribution procedure with integrated gradients. In primitive form, a path method explains a scalar output $F(\boldsymbol{x})$ using a *fiducial* image $\boldsymbol{x}^0$ as reference, which is presumably not associated with any class observed in training data. The importance attributed to *pixel* i for the purposes of explaining the quantity $F(\boldsymbol{x})$ is given by

$$\text{attr}_i^{\delta}(\boldsymbol{x}) = \int_0^1 \frac{\partial F(\delta(\alpha))}{\partial \delta_i(\alpha)} \frac{\partial \delta_i(\alpha)}{\partial \alpha} d\alpha$$

where $\delta : [0, 1] \to \mathbb{R}^n$ represents a curve with endpoints at $\delta(0) = \boldsymbol{x}^0$ and $\delta(1) = \boldsymbol{x}$. Here, $\sum_i \text{attr}_i(\boldsymbol{x}) = F(\boldsymbol{x}) - F(\boldsymbol{x}^0)$ follows from the *gradient theorem* for line integrals, s.t. the difference in output values decomposes over the sum of attributions. Commonly, $F(\boldsymbol{x}) = f_c(\boldsymbol{x})$ represents

the classification score for a class $c \in \mathcal{C}$ s.t. attributions capture elements in an image that are associated with this class. In order to attribute uncertainties, we readily assign $F(\boldsymbol{x}) = H(\boldsymbol{x})$, and thus combine scores across all classes with aims to identify pixels that *confuse* the model.

Integrated Gradients. Here, δ is parametrised as a straight path between a fiducial and the observed image, i.e. $\delta(\alpha) = \boldsymbol{x}^0 + \alpha(\boldsymbol{x} - \boldsymbol{x}^0)$, and the above simplifies to

$$\text{IG}_i(\boldsymbol{x}) = (x_i - x_i^0) \times \int_0^1 \frac{\partial H(\boldsymbol{x}^0 + \alpha(\boldsymbol{x} - \boldsymbol{x}^0))}{\partial x_i} d\alpha,$$

which corresponds to entropy attributions in Figure 1 (see Appendix A for its decomposition into aleatoric and epistemic attributions).

Integrated gradients offers an efficient approach to produce attributions with differentiable models, as an alternative to layer-wise relevance propagation [Montavon et al., 2019] or DeepLift [Shrikumar et al., 2017], and there exist several adaptations and extensions [Smilkov et al., 2017, Xu et al., 2020, Kapishnikov et al., 2021]. However, attributions are heavily influenced by differences in pixel values between $\boldsymbol{x}$ and $\boldsymbol{x}^0$, and the fiducial choice defaults to a black (or white) background. This fails to attribute importances to black (or white) pixels and is considered problematic [Sundararajan et al., 2017], leading to proposed *blurred* or *black+white* alternatives [Lundberg and Lee, 2017b, Kapishnikov et al., 2019]. Additionally, δ transitions the path $\boldsymbol{x}^0 \rightsquigarrow \boldsymbol{x}$ *out-of-distribution* [Jha et al., 2020, Adebayo et al., 2020], i.e. through intermediary images not representative of training data, leading to noise and artefacts in attributions.

3 METHODOLOGY

We describe the proposed method for uncertainty attributions summarised in Algorithm 1. This combines path integrals with a generative process to define a domain of integration. We use a counterfactual fiducial bearing no relation to causal inference [Pearl, 2010], i.e. an alternative *in distribution* image $\boldsymbol{x}^0$ similar to $\boldsymbol{x}$ according to a suitable metric, s.t. $f(\boldsymbol{x}^0)$ bears close to 0 predictive uncertainty.

We choose to leverage a *variational auto-encoder* (VAE) as the generative model. As customary, this is composed of a unit-Gaussian data-generating process of arbitrary dimensionality $m << n$, along with an image decoder $\psi : \mathbb{R}^m \to \mathbb{R}^n$. Here, $\boldsymbol{z}|\boldsymbol{x} \sim \mathcal{N}(\phi_\mu(\boldsymbol{x}), \phi_\sigma(\boldsymbol{x}))$ represents the approximate posterior in latent space, with mean and variance encoding functions $\phi_\mu, \phi_\sigma : \mathbb{R}^n \to \mathbb{R}^m$.

3.1 DOMAIN OF INTEGRATION

The domain of integration is defined as a curve across endpoints $\boldsymbol{x}^0 \rightsquigarrow \boldsymbol{x}$. We select the fiducial as a decoded image $\boldsymbol{x}^0 = \psi(\boldsymbol{z}^0)$, where $\boldsymbol{z}^0$ is the solution to the constrained optimization problem

$$\begin{aligned} \underset{\boldsymbol{z} \in \mathbb{R}^m}{\arg\min} \quad & \left[d(\psi(\boldsymbol{z}), \boldsymbol{x}) + \frac{1}{2m} \sum_j z_j^2 \right] \\ \text{subject to} \quad & \|e_{\hat{c}} - f(\psi(\boldsymbol{z}))\| < \varepsilon \end{aligned} \tag{2}$$

for an infinitesimal $\epsilon > 0$. Here, $\hat{c} = \arg\max_i f_i(\boldsymbol{x})$ is the predicted class by the classifier, and e_i is the unit indicator vector at index i. The metric $d(\cdot, \cdot)$ may be chosen to be the cross-entropy or mean absolute difference over pixel values in an image. The right-most term is the negative log-density (up to proportionality) of $\boldsymbol{z}$ in a latent space of dimensionality $m > 0$; this restricts the search *in-distribution* and ensures robustness to overparametrisation of the latent space within our experiments.

Hence, we retrieve a counterfactual fiducial which (i) is classified in the same class as $\boldsymbol{x}$ and (ii) bears close to zero predictive uncertainty. In practice, we approximate (2) through the penalty method, i.e. an unconstrained search with a large penalty on

$$d_{\mathcal{X}}(e_{\hat{c}}, f(\psi(\boldsymbol{z}))) = -\log f_{\hat{c}}(\psi(\boldsymbol{z})),$$

i.e. the cross-entropy between the predicted class $\hat{c}$ and the membership vector $f(\psi(\boldsymbol{z}))$ given a decoding $\psi(\boldsymbol{z})$. We proceed by gradient descent initialised at $\phi_\mu(\boldsymbol{x})$, the encoder's mean.

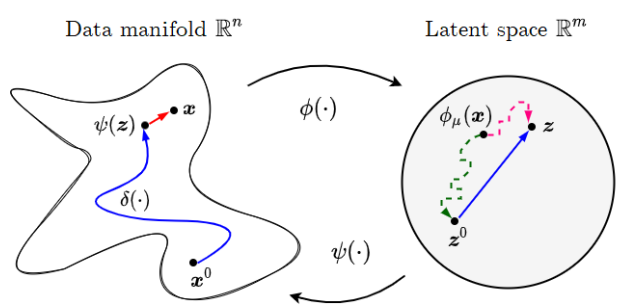

Figure 2: Procedural sketch to generate a path of integration. Here, *fiducial* $\boldsymbol{z}^0$ and *reconstruction* $\boldsymbol{z}$ points are optimized in latent space by gradient descent, starting initially from the encoding of $\boldsymbol{x}$ (dashed lines). A connecting straight path (in blue) is projected to the data-manifold and augmented with an interpolating component (in red).

Integration Path. We further leverage the decoder as a generative process to parametrise a curve $\delta_\psi : [0, 1] \to \mathbb{R}^n$, by following the steps displayed in Figure 2, s.t. $\delta_\psi(\alpha) = \psi(\boldsymbol{z}^0 + \alpha(\boldsymbol{z} - \boldsymbol{z}^0))$ where

$$\boldsymbol{z} = \underset{\boldsymbol{z} \in \mathbb{R}^m}{\arg\min} \left[d(\psi(\boldsymbol{z}), \boldsymbol{x}) + \frac{1}{2m} \sum_j z_j^2 \right]$$

is also optimised by gradient descent initialised at $\phi_\mu(\boldsymbol{x})$. This is an unconstrained optimisation problem analogue

Algorithm 1: Generative Uncertainty Attributions

input : Feature vector $\boldsymbol{x}$, predictive distribution $f(\cdot)$ and distance metric $d(\cdot,\cdot)$.
VAE encoder $\phi(\cdot)$ and decoder $\psi(\cdot)$, penalty $\lambda >> 0$ and learning rate $\nu > 0$.

output : Attributions $\text{attr}_i^{\delta_\psi}(\boldsymbol{x})$, $i = 1, \ldots, n$.

Initialise $\boldsymbol{z}^0 = \boldsymbol{z} = \phi_\mu(\boldsymbol{x})$;

Compute predicted class $\hat{c} = \arg\max_i f_i(\boldsymbol{x})$;

while $\mathcal{L}_1$ *not converged* **do**

$$\mathcal{L}_1 \leftarrow d(\psi(\boldsymbol{z}^0), \boldsymbol{x}) + \frac{1}{2m}\sum_j z_j^2 - \lambda \log f_{\hat{c}}(\psi(\boldsymbol{z})) \quad \text{and} \quad \boldsymbol{z}^0 \leftarrow \boldsymbol{z}^0 - \nu \nabla_{\boldsymbol{z}} \mathcal{L}_1$$

end

while $\mathcal{L}_2$ *not converged* **do**

$$\mathcal{L}_2 \leftarrow d(\psi(\boldsymbol{z}), \boldsymbol{x}) + \frac{1}{2m}\sum_j z_j^2 \quad \text{and} \quad \boldsymbol{z} \leftarrow \boldsymbol{z} - \nu \nabla_{\boldsymbol{z}} \mathcal{L}_2$$

end

Approximate $\text{attr}_i^{\delta_\psi}(\boldsymbol{x})$, $i = 1, \ldots, n$ in (3) along $\delta_{\psi, \boldsymbol{z}^0 \to \boldsymbol{z}}$ through trapezoidal integration.

to (2). Consequently, the path δ_ψ offers trajectory between a counterfactual $\delta_\psi(0) = \psi(\boldsymbol{z}^0) = \boldsymbol{x}^0$ and a reconstruction $\delta_\psi(1) = \psi(\boldsymbol{z})$ of the image $\boldsymbol{x}$. In order to correct for mild reconstruction errors, we finally augment the domain of integration through a *vanilla* straight path between the end-points $\psi(\boldsymbol{z}) \rightsquigarrow \boldsymbol{x}$. We display a few examples of this procedure on MNIST digits within Figure 3. Overall, the difference in predictive entropy or model scores between a reconstruction $\psi(\boldsymbol{z})$ and its original counterpart $\boldsymbol{x}$ are not observed to be significant within our experiments.

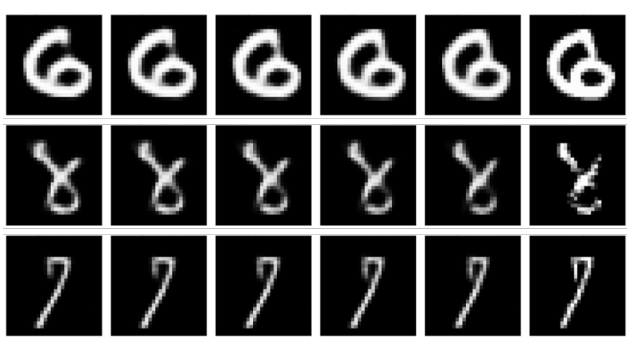

Figure 3: An example of *in-distribution* curves connecting fiducial (left-most) and real (right-most) data points, on MNIST digits data. Digits on the left bear no predictive uncertainty in classification.

3.2 LINE INTEGRAL FOR ATTRIBUTIONS

For simplicity, we restrict the formulae to the *in-distribution* component along the curve $\delta_\psi : [0,1] \to \mathbb{R}^n$ defined in Subsection 3.1, and we ignore the straight path connecting $\psi(\boldsymbol{z}) \rightsquigarrow \boldsymbol{x}$. We require the total differential of the entropy $H(\cdot)$ wrt $\boldsymbol{z}$ in latent space; however, we wish to retrieve importances for features $\boldsymbol{x}$ in the original data manifold within $\mathbb{R}^n$. To this end, the attribution at index $i = 1, \ldots, n$ is given by

$$\text{attr}_i^{\delta_\psi}(\boldsymbol{x}) = \sum_{j=1}^{m}(z_j - z_j^0)\int_0^1 \frac{\partial H(\delta_\psi(\alpha))}{\partial \delta_{\psi,i}(\alpha)} \frac{\partial \delta_{\psi,i}(\alpha)}{\partial z_j} d\alpha. \tag{3}$$

Intuitively, we compute the total derivative of $H(\cdot)$ wrt α in the integration path, using the chain rule. We decompose the calculation over indices in pixel space, and further undertake summation over contributions in latent space. In Figure 4, we show an example that compares attributions in (3) versus vanilla integrated gradients. There, we find a *CelebA* image [Liu et al., 2015] with tags for the presence of a *smile*, *arched eyebrows* and *no bags under the eyes*.

3.3 PROPERTIES

Due to *path independence* and noting that $H(\boldsymbol{x}^0) \approx 0$ by definition, importances drawn from a trajectory $\delta_\psi(\cdot)$ as parametrised in Subsection 3.1 will approximately account for **all** of the uncertainty in a posterior predictive task, i.e.

$$H(\boldsymbol{x}) \approx \int_0^1 \nabla H(\delta_\psi(\alpha)) d\alpha = \sum_{i=1}^{n} \text{attr}_i^{\delta_\psi}(\boldsymbol{x}),$$

and this is commonly referred to as *completeness*. Additionally, the reliance on path derivatives along with the rules of composite functions ensure that both fundamental axioms of *sensitivity(b)* (i.e. *dummy property*) along with *implementation invariance* are inherited, and we refer the reader to Friedman [2004], Sundararajan et al. [2017] for the technical details. Importantly, the attribution will be zero for any index which does not influence the classifier.

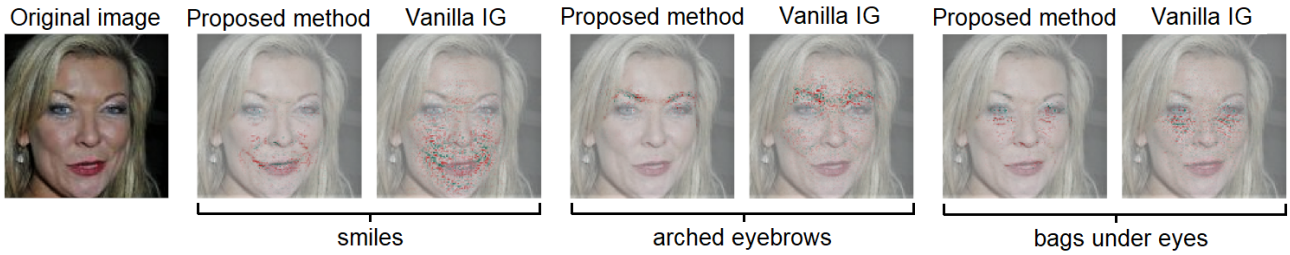

Figure 4: Comparison of uncertainty attributions on a *CelebA* image. We compare attributions for three classifiers, which measure the presence (or lack) of *smiles* (left), *arched eyebrows* (centre), and *bags under eyes* (right). Red pixels contribute by increasing uncertainties, in green we find contributions towards decreasing uncertainties.

3.3.1 The Role of the Autoencoder

A VAE is arguably not the best generative model for reconstructing sharp images with high fidelity. However, it is stable during training and efficient in sampling, furthermore, the encoder provides a mean to efficiently select starting values $\phi_\mu(\boldsymbol{x})$ during latent optimisation tasks [cf. Antoran et al., 2021]. In Appendix B we offer a robustness assessment of our results to variations in the autoencoder, and we report on negligible changes in performance. We achieve consistency even in large overparametrised latent spaces, due to Gaussian priors in the optimisation procedures in Subsection 3.1, which define the integration path.

Alternative models can be used to define integration paths. *Generative adversarial networks* have gained relevance as a means to facilitate interpretability in classification tasks [Lang et al., 2021], however, training can be unstable and identifying counterfactual references is infeasible. This also presents a problem with *autoregressive models* [Van den Oord et al., 2016], which are further inefficient in sampling and would pose long optimisation times in latent space.

3.3.2 Non-Generative Integration Paths

For simplicity, a counterfactual fiducial image $\boldsymbol{x}^0 = \psi(\boldsymbol{z}^0)$ as described in (2) can also be combined with a straight or *guided* [Kapishnikov et al., 2021] integration path $\psi(\boldsymbol{z}^0) \rightsquigarrow \boldsymbol{x}$. In application to simple grey-scale images, this path is unlikely to transverse *out-of-distribution* due to the proximity between a fiducial and the original image $\boldsymbol{x}$. In our experiments, we test these variants and report that they fare relatively well in explainability tasks with simple images; however, their performance degrades on complex RGB pictures involving facial features.

4 EXPERIMENTS

We apply our proposed methodology to classification models for benchmark image data sets of increasing complexity, including the repositories *MNIST handwritten digits* [LeCun and Cortes, 2010], *fashion-MNIST* [Xiao et al., 2017] and *CelebA* [Liu et al., 2015]. We evaluate the performance both quantitatively and qualitatively, and we compare the results to path methods including *vanilla* integrated gradients [Sundararajan et al., 2017], as well as *blur* and *guided* variants [Xu et al., 2020, Kapishnikov et al., 2021]. We test these approaches with *plain*, *black+white* (B+W) and *counterfactual* fiducials, and we combine the saliency maps with Xrai [Kapishnikov et al., 2019], a popular segmentation and attribution approach. We also evaluate pure counterfactual approaches for uncertainty attributions, which assign importances by directly comparing pixel values between an image and its counterfactual. For this, we include most recent *CLUE* attributions [Antoran et al., 2021] in the assessment. For completeness, we finally add adaptations of *LIME* [Ribeiro et al., 2016] and *kernelSHAP* [Lundberg and Lee, 2017a]. Implementation details can be found in Appendix D. Source code for reproducing these results can be found at *[removed for blind review]*.

4.1 PERFORMANCE METRICS

In order to produce quantitative evaluations we resort to *smallest sufficient region* methods popularised in recent literature [see Petsiuk et al., 2018, Kapishnikov et al., 2019, Covert et al., 2020, Lundberg et al., 2020], which evaluate the quality of saliency maps in the absence of ground truths. These are suitable for our repeated assessments over multiple methods and data sets, as they do not require for specialised model retrains [cf. Hooker et al., 2019, Jethani et al., 2021]. The methods proceed by revealing pixels from a masked image, in order of importance as determined by attribution values, and changes in classification scores, predictive entropy or image information content are monitored. Alternatively, the process may be carried backwards by removing or resampling pixels from the original image, and we show an example of this process in Figure 5. We use blurring as a masking mechanism [cf. Kapishnikov et al., 2019], since other alternatives lead to masked images significantly out of distribution, i.e. non representative of training data. We evaluate two inclusion and removal metrics suitable to measure changes in predictive uncertainty.

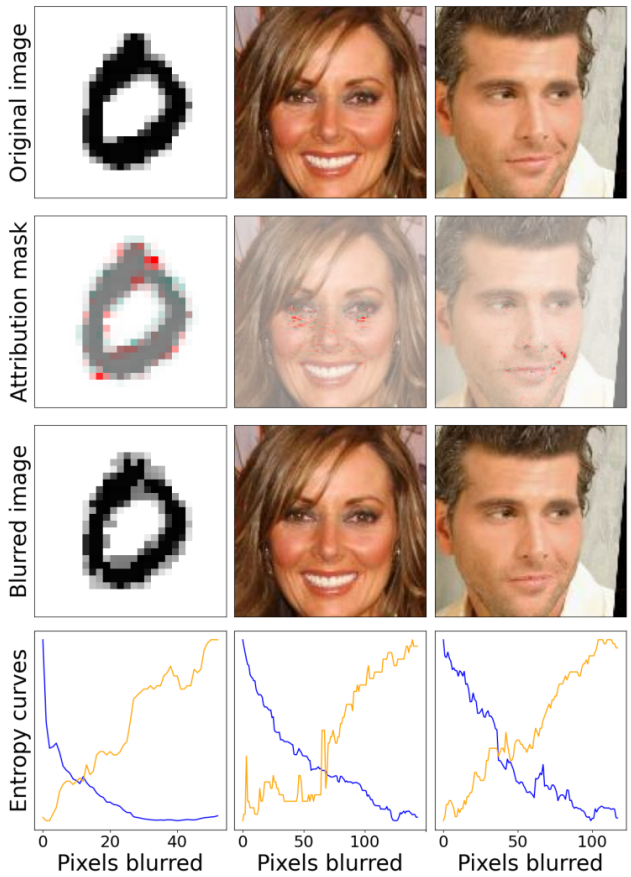

Figure 5: Normalised variation in predictive entropy (decreasing, blue) and image information content (increasing, orange) as pixels most contributing to uncertainty are sequentially blurred. Classification task on digits (left), bags under eyes (centre) and smiles (right). Information content approximated by compressed file sizes.

Inclusion Methods. We measure the *entropy information curve* (EIC) in a manner analogue to *performance information curves* (PICs) discussed in Kapishnikov et al. [2019, 2021]. For an image $\boldsymbol{x}$ with n pixels, we define a sequence $\{\boldsymbol{x}^i\}_{i=0,\dots,n}$ that transitions from a blurred reference $\boldsymbol{x}^0 = \boldsymbol{x}_{blurred}$ towards $\boldsymbol{x}^n = \boldsymbol{x}$, by revealing pixels in order of contribution to decreasing the entropy. We evaluate

$$EIC_i = \frac{1}{|\mathcal{X}|} \sum_{\boldsymbol{x} \in \mathcal{X}} \frac{H(\boldsymbol{x}^i)}{H(\boldsymbol{x}_{blurred})}$$

across indexes in the transition $\boldsymbol{x}_{blurred} \xrightarrow{i=1,\dots,n} \boldsymbol{x}$, which retrieves an average over images in each data set $\mathcal{X}$ (in the presence of significant outliers, we report on median values). The EIC measures the variation in overall predictive entropy and can be computed on unlabelled data. It is assessed versus the information content in the images as pixels are revealed Kapishnikov et al. [2019, 2021], which can be approximated by file sizes or the second order Shannon entropy.

Best Removal Methods. We measure *uncertainty reduction curves*, i.e. the relative uncertainty that an attribution method can remove from an image $\boldsymbol{x}$. We use the inverse sequence $\{\boldsymbol{x}^i\}_{i=0,\dots,n}$, which transitions from $\boldsymbol{x}^0 = \boldsymbol{x}$ towards a *blurred* image $\boldsymbol{x}^n = \boldsymbol{x}_{blurred}$. We evaluate

$$URC_i = \frac{1}{|\mathcal{X}|} \sum_{\boldsymbol{x} \in \mathcal{X}} max_{r \leq i} \left[1 - \frac{H(\boldsymbol{x}^r)}{H(\boldsymbol{x})} \right],$$

i.e. the best percentage reduction in predictive uncertainty that can be explained away by *blurring* up to i pixels, in decreasing order of contribution to uncertainty.

4.2 QUANTITATIVE EVALUATION

In Table 1 we report on (i) the area over the entropy information curve and (ii) percentile points in the uncertainty reduction curve, for the various attribution methods and data sets analysed in this paper. We explore 5 classification tasks, including the presence of *smiles*, *arched eyebrows* and *eye-bags* in CelebA images. In all cases, high values represent better estimated performance. The metrics are evaluated on images that were excluded during model training. Attribution methods have been implemented with default parameters, where available, and we offer details in Appendix D. Blurring is performed with a Gaussian kernel, and the standard deviation is tuned individually for each classification task. We choose the minimum standard deviation s.t. a model's predictive uncertainty for the fully blurred images is maximised. *KernelSHAP* evaluations are offered only for data sets with small resolution images, due to the computational complexity associated with undertaking the recommended amount of image perturbations.

The results show that a generative method as presented in this paper is better suited to explain variations in predictive entropy, as well as explaining away sources of uncertainty. Results suggest that improvements over the explored alternatives are of significance in classification tasks with high resolution images concerning facial features. In application to low resolution grey scale images, the results also show that popular attribution approaches, such as integrated gradients, guided integrated gradients or SHAP require a counterfactual fiducial to perform well, which must still be produced through a generative model. In these cases, good performance is a consequence of low dissimilarity between an image and its baseline (see Subsection 3.3.2), s.t. simple integration paths remain in-distribution.

Blurring setting. Evaluations are notoriously dependent on the standard deviation setting of the Gaussian kernel. High standard deviation settings lead to blurred images that are significantly out of distribution. This degrades the projected performance across all attribution methods, as observed in the URC curves displayed in Figure 6, corresponding to the classification model for *bags under the eyes* on CelebA data.

Table 1: Area over the entropy information curve and percentile points in uncertainty reduction curves, across attribution methods and classification tasks. Metrics procured wrt approximated and normalised information content of images.

Method	Area over Entropy Information Curve					Uncertainty Reduction Curve									
	Mnist	Fashion	Smiles	Eyebrows	Eyebags	Mnist		Fashion		Smiles		Eyebrows		Eyebags	
						1%	5%	1%	5%	5%	10%	5%	10%	5%	10%
Vanilla IG	0.998	0.759	0.354	0.155	0.143	0.469	0.508	0.109	0.196	0.076	0.085	0.097	0.104	0.117	0.131
+ (B+W)	**0.999**	0.901	0.584	0.422	0.361	0.379	0.631	0.083	0.217	0.149	0.185	0.209	0.233	0.146	0.195
+ Counterfactual	**0.999**	0.909	0.600	0.396	0.325	**0.751**	**0.872**	**0.217**	**0.431**	0.176	0.215	0.213	0.244	0.153	0.179
Blur IG	0.973	0.818	0.368	0.144	0.136	0.017	0.102	0.016	0.076	0.015	0.019	0.014	0.017	0.008	0.009
Guided IG	0.996	0.655	0.333	0.134	0.119	0.222	0.291	0.009	0.035	0.014	0.017	0.016	0.023	0.009	0.012
+ (B+W)	0.997	0.735	0.318	0.151	0.130	0.115	0.283	0.006	0.036	0.017	0.018	0.035	0.046	0.008	0.013
+ Counterfactual	**0.999**	0.879	0.360	0.277	0.206	0.715	0.833	0.168	0.326	0.056	0.063	0.137	0.152	0.062	0.081
Generative IG	**0.999**	**0.920**	**0.737**	**0.429**	**0.433**	0.747	0.866	0.201	0.386	**0.318**	**0.389**	**0.243**	**0.278**	**0.173**	**0.233**
LIME	0.993	0.630	0.231	0.088	0.140	0.000	0.021	0.001	0.011	0.011	0.015	0.009	0.016	0.009	0.019
SHAP	0.994	0.900				0.119	0.319	0.080	0.222						
+ Counterfactual	0.985	0.839				0.515	0.683	0.165	0.302						
CLUE	0.969	0.659	0.349	0.177	0.135	0.264	0.289	0.042	0.076	0.028	0.031	0.043	0.050	0.007	0.010
XRAI + IG	0.991	0.750	0.541	0.230	0.156	0.023	0.093	0.010	0.037	0.053	0.101	0.036	0.056	0.018	0.028
+ (B+W)	0.992	0.811	0.637	0.312	0.236	0.002	0.035	0.009	0.044	0.121	0.206	0.067	0.103	0.028	0.057
+ Counterfactual	0.952	0.648	0.267	0.235	0.243	0.248	0.425	0.057	0.148	0.098	0.144	0.134	0.227	0.102	0.183
XRAI + GIG	0.990	0.671	0.173	0.098	0.054	0.012	0.054	0.003	0.016	0.019	0.030	0.006	0.012	0.003	0.005
+ (B+W)	0.988	0.699	0.118	0.120	0.043	0.001	0.018	0.002	0.012	0.021	0.032	0.016	0.027	0.002	0.004
+ Counterfactual	0.960	0.622	0.094	0.222	0.115	0.202	0.391	0.028	0.087	0.012	0.013	0.082	0.107	0.010	0.019
XRAI + Gen IG	0.971	0.710	0.512	0.240	0.275	0.245	0.415	0.047	0.129	0.179	0.243	0.141	0.224	0.113	0.190

Thus, results in Table 1 represent *best* measured performances. Also, we note that attributions produced in combination with Xrai [Kapishnikov et al., 2019] remain consistent across evaluations, a benefit from pre-processing and pixel segmentation leading to highly clustered importances.

Autoencoder Settings. The performance of our proposed method plateaus after a certain dimensionality is reached in the latent space representation. Further increasing the complexity of the autoencoder, or changing its training scheme, leads to consistent results. This is a consequence of regularisation terms imposed over optimisation tasks in (2). We note that fiducial points and integration paths are forced to lie in distribution, even within large and overparametrised encoding spaces. A robustness assessment with performance metrics can be found in Appendix B.

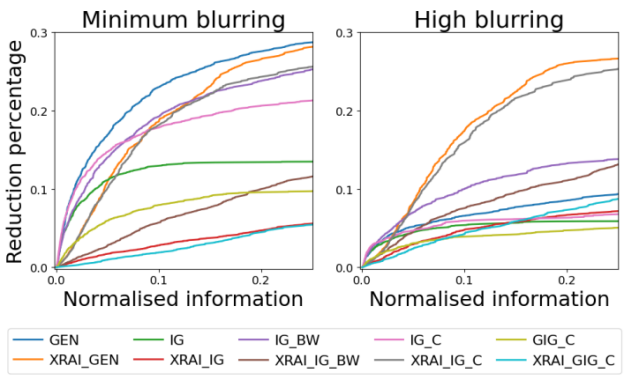

Figure 6: Uncertainty reduction curves for best performing attribution methods on *bags under the eyes*, CelebA data. Left, blurring is set to the minimum feasible value. Right, we assign an arbitrarily large standard deviation.

4.3 QUALITATIVE EVALUATION

In Figure 7 we find sample uncertainty attribution masks associated with best performing methods, and we offer further examples in Appendix C. In the figure, attribution masks for vanilla IG and guided IG are presented with counterfactual fiducial baselines, in order to avoid noisy saliency masks, such as previously observed in Figure 4. Counterfactual baselines allow to isolate small subsets of pixels that are associated with predictive uncertainty, and producing them requires an autoencoder. In combination with an integration path further defined by a generative model, the attribution method we have presented produces clustered attributions which are de-correlated from raw pixel-value differences between an image and its counterfactual, unlike *Clue* importances. This offers increasingly sparse and easily interpretable uncertainty attributions, which is reportedly associated with better performance in quantitative evaluations [cf. Kapishnikov et al., 2021]. Finally, segmentation based

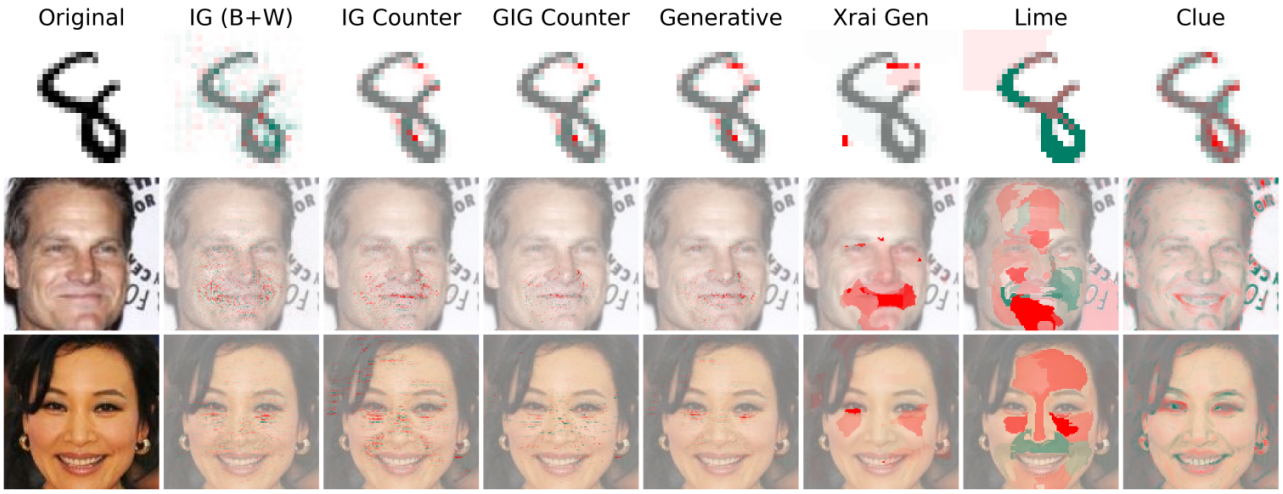

Figure 7: Sample uncertainty attribution masks for selected attribution methods. Masks correspond to digits (top), smiles (mid) and bags under the eyes (bottom).

mechanisms do not perform well in the data sets that we have explored, since they do not contain varied objects and items that can be easily segregated.

5 DISCUSSION

In this paper, we have introduced a novel framework for the attribution of predictive uncertainties in classification models, which combines path methods, counterfactual explanations and generative models. This is thus an additional tool contributing to improved transparency and interpretability in deep learning applications.

We have further offered comprehensive benchmarks on the multiple approaches for explaining predictive uncertainties, as well as vanilla adaptations of popular score attribution methods. For this purpose, we have leveraged standard feature removal and addition techniques. Our experimental results show that a combination of counterfactual fiducials along with straight or guided path integrals is sufficient to attain best performance in simple classification tasks with greyscale images. However, complex images benefit from subtle definitions of integration paths that can only be defined through a generative process as described in this paper.

The method presented in this paper is applicable to classification models for data sets where we may feasibly synthesise realistic images through a generative model. This currently includes a variety of application domains, such as human faces, postures, pets, handwriting, clothes, or landscapes [Creswell et al., 2018]. Yet, the scope and ability of such models to synthesise new types of figures is quickly increasing. Also, we evidenced that we do not require a particularly accurate generative process within our method, i.e. the uncertainty attribution procedure we have presented yields top performing results even in the presence of errors and dissimilarities during image reconstructions.

References

Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. In *Advances in Neural Information Processing Systems*, volume 33, pages 700–712, 2020.

Javier Antoran, Umang Bhatt, Tameem Adel, Adrian Weller, and José Miguel Hernández-Lobato. Getting a CLUE: A method for explaining uncertainty estimates. In *International Conference on Learning Representations*, 2021.

Pranjal Awasthi, Alex Beutel, Matthäus Kleindessner, Jamie Morgenstern, and Xuezhi Wang. Evaluating fairness of machine learning models under uncertain and incomplete information. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 206–214, 2021.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622, 2015.

Ian Covert, Scott Lundberg, and Su-In Lee. Feature removal is a unifying principle for model explanation methods. *arXiv preprint arXiv:2011.03623*, 2020.

Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65, 2018.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pages 1184–1193. PMLR, 2018.

Eric J Friedman. Paths and consistency in additive cost sharing. *International Journal of Game Theory*, 32(4): 501–518, 2004.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 2019.

Xianxu Hou, Linlin Shen, Ke Sun, and Guoping Qiu. Deep feature consistent variational autoencoder. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1133–1141, 2017. doi: 10.1109/WACV.2017.131.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3): 457–506, 2021.

Neil Jethani, Mukund Sudarshan, Yindalon Aphinyanaphongs, and Rajesh Ranganath. Have we learned to explain?: How interpretability methods can learn to encode predictions in their interpretations. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Anupama Jha, Joseph K Aicher, Matthew R Gazzara, Deependra Singh, and Yoseph Barash. Enhanced integrated gradients: improving interpretability of deep learning models using splicing codes as a case study. *Genome biology*, 21(1):1–22, 2020.

Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai: Better attributions through regions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4948–4957, 2019.

Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. Guided integrated gradients: An adaptive path method for removing noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5050–5058, 2021.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip Isola, Amir Globerson, Michal Irani, et al. Explaining in style: Training a gan to explain a classifier in stylespace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 693–702, 2021.

Yann LeCun and Corinna Cortes. MNIST handwritten digits database, 2010. `http://yann.lecun.com/exdb/mnist/`.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 4768–4777, 2017a.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017b.

Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209, 2019.

Judea Pearl. Causal inference. In *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008*, pages 39–58, 2010.

Iker Perez and Giuliano Casale. Variational inference for markovian queueing networks. *Advances in Applied Probability*, 53(3):687–715, 2021.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Lisa Schut, Oscar Key, Rory Mc Grath, Luca Costabello, Bogdan Sacaleanu, Yarin Gal, et al. Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleatoric uncertainties. In *International Conference on Artificial Intelligence and Statistics*, pages 1756–1764. PMLR, 2021.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153, 2017.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, 2017.

Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI*, 2018.

Jost Tobias Springenberg, Aaron Klein, Stefan Falkner, and Frank Hutter. Bayesian optimization with robust bayesian neural networks. *Advances in neural information processing systems*, 29:4134–4142, 2016.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15:1929–1958, 2014.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

Omer Faruk Tuna, Ferhat Ozgur Catak, and M Taner Eskil. Exploiting epistemic uncertainty of the deep learning models to generate adversarial samples. *arXiv preprint arXiv:2102.04150*, 2021.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, pages 9690–9700. PMLR, 2020.

Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.

Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. *arXiv preprint arXiv:1907.02584*, 2019.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. Data available at `https://github.com/zalandoresearch/fashion-mnist`.

Yijun Xiao and William Yang Wang. Quantifying uncertainties in natural language processing tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7322–7329, 2019.

Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. Attribution in scale and space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9680–9689, 2020.

A BAYESIAN PRESENTATION

In Bayesian settings, model uncertainties are often decomposed across *aleatoric* and *epistemic* components that help scrutinise different aspects in the functioning of a model, and can facilitate interpretability or fairness assessments in important machine learning applications [Awasthi et al., 2021]. Hence we may wish to offer attributions that are representative of isolated types of uncertainties.

On training a neural classifier $f : \mathbb{R}^n \times \mathcal{W} \to \Delta^{|\mathcal{C}|-1}$ within an (approximate) Bayesian setting, we commonly obtain a *posterior* over the hypothesis space of models, i.e. a distribution $\pi(\boldsymbol{w}|\mathcal{D})$ over model weights conditioned on the available train data $\mathcal{D} = \{\boldsymbol{x}_i, c_i\}_{i=1,2,\ldots}$. Popular approaches to procure such posterior often differ in their approach to incorporate *prior* knowledge and include *dropout* [Srivastava et al., 2014], *Bayes-by-Backprop* [Blundell et al., 2015] or SG-HMC [Springenberg et al., 2016]. Here, a model score for a new data point $\boldsymbol{x}^\star \in \mathbb{R}^n$ is derived from the *posterior predictive distribution* by marginalising over posterior weights, i.e.

$$\pi(\boldsymbol{x}^\star|\mathcal{D}) = \int_{\mathcal{W}} f(\boldsymbol{x}^\star, \boldsymbol{w})\pi(\boldsymbol{w}|\mathcal{D})d\boldsymbol{w} = \mathbb{E}_{\boldsymbol{w}|\mathcal{D}}[f(\boldsymbol{x}^\star, \boldsymbol{w})],$$

and is easily approximated as $\frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{x}^\star, \boldsymbol{w}_i)$, with weight samples $\boldsymbol{w_i} \sim \pi(\boldsymbol{w}|\mathcal{D})$, $i = 1, \ldots, N$. This setting is analogue to the presentation in Section 2, however, the point estimate score must now be averaged over the posterior, i.e. $f(\boldsymbol{x}) = \pi(\boldsymbol{x}^\star|\mathcal{D}) = \mathbb{E}_{\boldsymbol{w}|\mathcal{D}}[f(\boldsymbol{x}^\star, \boldsymbol{w})]$

The *entropy* is thus given by

$$H(\boldsymbol{x}|\mathcal{D}) = -\sum_{c\in\mathcal{C}} \mathbb{E}_{\boldsymbol{w}|\mathcal{D}}[f_c(\boldsymbol{x}, \boldsymbol{w})] \cdot \log \mathbb{E}_{\boldsymbol{w}|\mathcal{D}}[f_c(\boldsymbol{x}, \boldsymbol{w})],$$

and may be decomposed through the law of iterated variances [Kendall and Gal, 2017] so as to yield an *aleatoric* term

$$\begin{aligned} H_a(\boldsymbol{x}|\mathcal{D}) &= \mathbb{E}_{\boldsymbol{w}|\mathcal{D}}[H(\boldsymbol{x}, \boldsymbol{w})] \\ &= -\sum_{c\in\mathcal{C}} \mathbb{E}_{\boldsymbol{w}|\mathcal{D}}\big[f_c(\boldsymbol{x}, \boldsymbol{w}) \cdot \log f_c(\boldsymbol{x}, \boldsymbol{w})\big], \end{aligned}$$

which measures the mean predictive entropy across models in the posterior hypothesis space, as well as the *mutual information* or *epistemic* term, $H_e(x|\mathcal{D}) = H(\boldsymbol{x}|\mathcal{D}) - H_a(\boldsymbol{x}|\mathcal{D})$ that represents model uncertainty projected into the latent membership vector $\pi(\boldsymbol{x}|\mathcal{D})$. Intuitively, aleatoric uncertainty represents natural stochastic variation in the observations over repeated experiments; on the other hand, epistemic uncertainty is descriptive of model unknowns due to inadequate data or inappropriate modelling choices.

Path integrals. The *posterior predictive classifier* $\pi(\boldsymbol{x}|\mathcal{D})$ accepts a path importance for an arbitrary scalar output $F(\boldsymbol{x}, \boldsymbol{w})$ at index i, given by

$$\text{attr}_i^\delta(\boldsymbol{x}) = \int_0^1 \mathbb{E}_{\boldsymbol{w}|\mathcal{D}}\left[\frac{\partial F(\delta(\alpha), \boldsymbol{w})}{\partial \delta_i(\alpha)}\right] \frac{\partial \delta_i(\alpha)}{\partial \alpha} d\alpha.$$

This represents a *mean-average* trajectory over a curve δ and follows from *dominated convergence*. This easily amends to the attribution of uncertainties, i.e.

$$\text{attr}_i^\delta(\boldsymbol{x}) = -\sum_{c\in\mathcal{C}} \int_0^1 \Delta_i(\alpha) \frac{\partial \delta_i(\alpha)}{\partial \alpha} d\alpha$$

which is defined s.t.

$$\Delta_i(\alpha) = \big(1 + \log \mathbb{E}_{\boldsymbol{w}|\mathcal{D}}[f_c(\delta(\alpha), \boldsymbol{w})]\big) \cdot \mathbb{E}_{\boldsymbol{w}|\mathcal{D}}\Big[\frac{\partial f_c(\delta(\alpha), \boldsymbol{w})}{\partial \delta_i(\alpha)}\Big]$$

If we wish to only attribute aleatoric uncertainties, we may replace the above for

$$\Delta_i(\alpha) = \mathbb{E}_{\boldsymbol{w}|\mathcal{D}}\Big[\big(1 + \log f_c(\delta(\alpha), \boldsymbol{w})\big) \cdot \frac{\partial f_c(\delta(\alpha), \boldsymbol{w})}{\partial \delta_i(\alpha)}\Big].$$

Finally, attributions for any variation in epistemic uncertainty is readily shown to be *explained* as the difference in attributions between full and aleatoric uncertainties. We showed an example of this in Figure 1 within Section 2.

B ROBUSTNESS TO CHANGES IN THE AUTOENCODER

In Table 2 we show evaluations of performance metrics for the attribution method proposed in this paper, over resampled Mnist and Fashion validation images. We train multiple variational autoencoders and use them as the generative process to define integration paths in our method. These differ in the dimensionality of the latent space used to encode reduced representations of images. This is the most impactful layer for the functioning of the attribution method we have presented, since straight integration lines are defined in this space and later projected into pixel space. Too small or large a space could lead to out of distribution images and

Table 2: Performance metrics for generative attribution method, for architecture variations in the autoencoder.

Setting	Area over EIC		Uncertainty Reduction Curve			
	Mnist	Fashion	Mnist		Fashion	
			1%	5%	1%	5%
4	0.999	0.918	0.474	0.738	0.165	0.350
8	0.999	0.916	0.661	0.845	0.192	0.374
16	0.999	0.919	0.704	0.846	0.196	0.393
32	0.999	0.925	0.743	0.868	0.204	0.395
- Aug	0.999	0.930	0.687	0.876	0.184	0.403
64	0.999	0.922	0.752	0.877	0.203	0.392
128	0.999	0.925	0.756	0.879	0.206	0.400
259	0.999	0.927	0.762	0.884	0.204	0.405

integration paths. Additionally, we also experiment with altering the data augmentation mechanism used for modifying images prior to training the autoencoder (results are reported at latent space dimension of 32). No significant changes in performance where noticed as training regimes and learning rates were modified.

In the table, we notice consistent performance which plateaus after a certain threshold, which is equivalent in these two data sets. Consistency in performance is a consequence of the regularisation term in latent space observed in (2). This tunes fiducial points and integration paths strictly in distribution, even if large latent spaces overparametrise the encoding space.

C EXAMPLES

In Figure 8 we find examples of attributions of aleatoric and epistemic uncertainty types, applied to dog versus cats images. Attributions are produced by vanilla integrated gradients as described in Section 2. Saliency masks are combined with a Gaussian kernel in order to draw attention to regions in images associated with different uncertainty types. Similarly, Figure 9 shows attributions of uncertainty types across selected Mnist images, produced by the generative method presented in this paper.

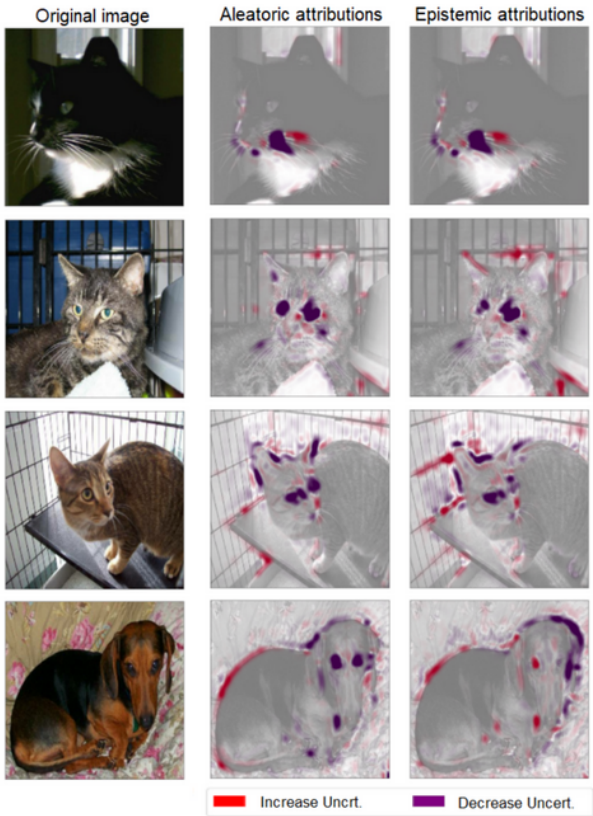

Figure 8: Aleatoric and epistemic contributions to uncertainty for a classification task in *dogs versus cats* data.

Figure 9: Aleatoric and epistemic contributions to uncertainty for a classification task with MNIST digits.

C.1 QUALITATIVE EVALUATIONS

Finally, in Figure 10 we show uncertainty attribution masks across a range of classification tasks, on all the data sets explored in this paper. In all cases, we note that attributions relying on counterfactual mechanisms are humanly interpretable. Further integration of counterfactual methods with path integrals ensures that attributions are isolated to few pixels. In application to human gestures, these are always restricted to facial features around the mouth, cheeks or eyebrows, depending on the classification task. On the contrary, vanilla attributions through integrated gradients (averaged over *black and white* fiducial baselines) are noticeably noisy. Also, segmentation based mechanisms do not perform well in the data sets we have explored, which do not contain multiple objects that can be easily segregated.

D IMPLEMENTATION DETAILS

All of our predictive models are implemented through Keras. The following is a summary of architectures, hyperparameters, training regimes and further details.

D.1 MNIST HANDWRITTEN DIGITS

Our classifier is a convolutional neural network with *max-pooling* layers and dropout, structured as:

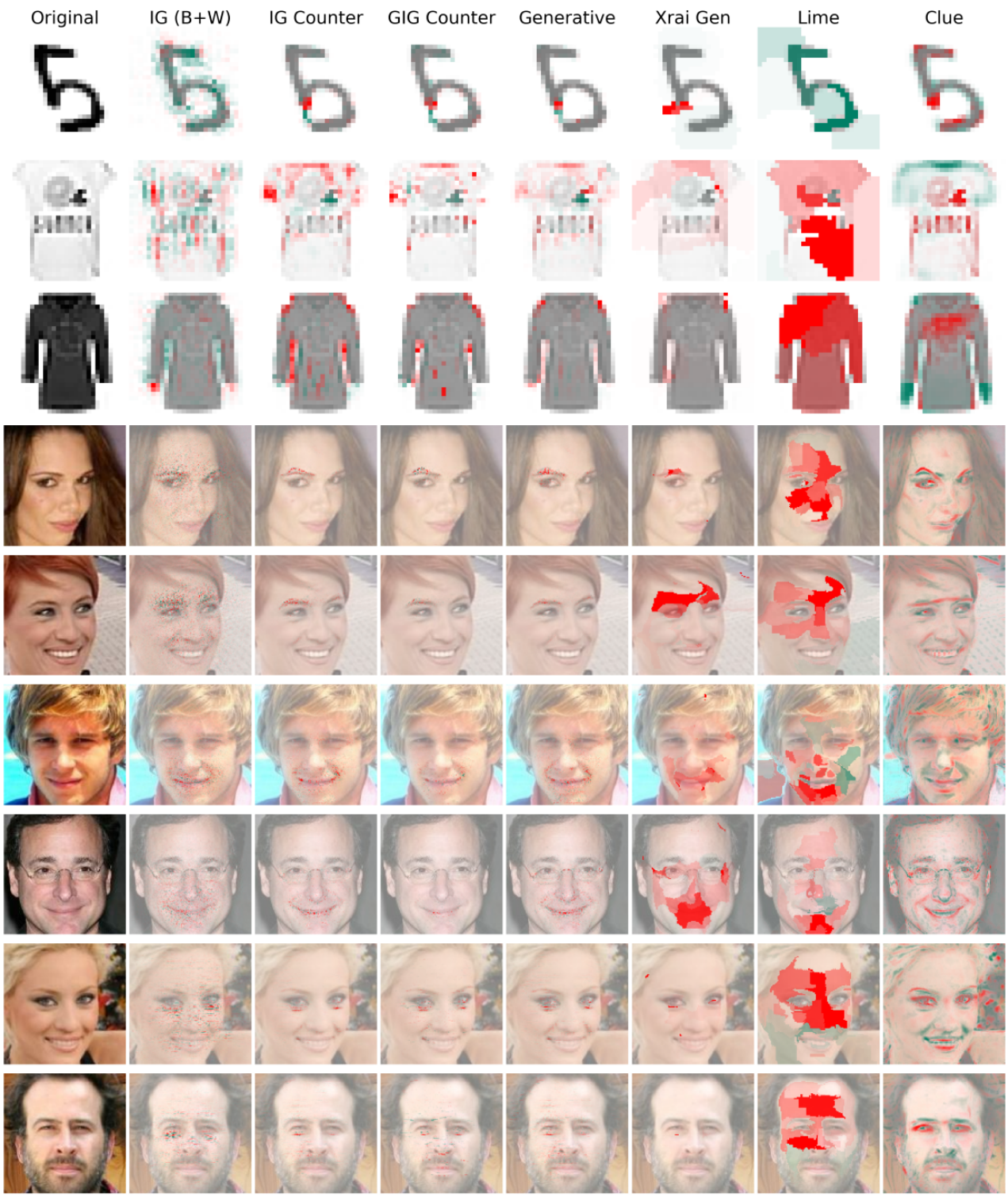

Figure 10: Uncertainty attribution masks across multiple classification tasks and data sets. We display best performing attribution methods with a counterfactual mechanism.

- Two convolutional layers of kernel size 3×3 and *relu* activation; *filter counts* are 32 and 64 for the first and second layers. We use *stride* length of 1 followed by *max-pooling* layers of *pool size* 2×2.
- The output is flattened and fed through a *dense* layer of 128 neurons with *relu* activation, followed by dro-

pout with deactivation rate of 0.5, and a final *softmax* regression layer for categorical outputs.

We train to minimize the *categorical cross entropy* wrt the train labels, using the *Adam* optimizer, over 10 epochs, with a constant learning rate of $1e^{-3}$ and with *batch size* of 32.

The **variational autoencoder** relies on convolution and deconvolution layers. The encoder is structured as:

- Two convolutional layers of kernel size 3×3, stride 2 and *relu* activation; *filter counts* are 32 and 64 for the first and second layers.
- A *dense* layer of 128 neurons, with *relu* activation.
- Two *dense* layers mapping the 128 neurons to a distributional mean vector and a log-standard-deviation vector, for the latent space for an image. Dimension of the latent space varies in order to assess robustness, see Appendix B for details.
- A random *sampling* operation from a normal distribution, with the afore-defined distributional parameters.

In addition, the decoder is defined as:

- A dense layer with *relu* activation, mapping a latent element to a vector of dimensionality $7 \times 7 \times 64$.
- Two deconvolutional layers of kernel size 3×3, stride length 2 and *relu* activation; the *filter counts* are 64 and 32 for the first and second layers.
- An output deconvolutional layer of kernel size 3×3, *filter counts* 1, stride length 1 and *sigmoid* activation for pixel values.

The autoencoder is fitted to minimize a custom loss, with a reconstruction term (through a cross-entropy loss) and the Kullback-Leibler divergence among latent mappings and a normal distribution $\mathcal{N}(\mathbf{0}, I)$. We use the *Adam* optimizer, over 50 epochs, with a constant learning rate of $1e^{-3}$ and with *batch size* of 32.

D.2 FASHION-MNIST DATASET

The classifier and autoencoder are defined similarly to the above example. However, we add two additional *dropout* layers (with probability 0.5) after each *max-pooling* operation in the classifier. Training proceeds with the *Adam* optimizer, at a constant learning rate of $1e^{-3}$ with *batch size* 32. The classifier is trained for 10 epochs using the cross-entropy as the cost function. The autoencoder is trained for 50 epochs using a combination of binary cross-entropy and the Kullback-Leibler divergence as a regularisation term.

D.3 CELEBA DATASET

Images are centred around the face and cropped to size 128×128, further standardized to pixel values in the range $[0, 1]$. During training, we leverage data augmentation with random rotations; we use a *maximum angle* of ± 18 degrees, random translation by a maximum factor of 0.1 and random horizontal flip.

The **classifier** is composed of 6 convolutional blocks followed by a dense layer with *softmax* activation. Each convolutional block utilizes a *kernel* size of 3 and *stride* 1, along with *batch normalization*, *dropout* with deactivation probability of 0.2, *relu* activation and *max-pooling* (*pool size* 2 and *stride* 2). The number of channels in convolutional layers is, respectively, 32, 64, 128, 128, 256 and 256. The last block is followed by a flattening operation and a *dropout* layer with deactivation probability 0.4.

We train this classifier for 5 epochs using the *Adam* optimizer with batch size 64 and the *cross-entropy* as cost function. The learning rate is decreased after each epoch by a factor of 0.8; starting from $1e^{-4}$ for the *smiling* and *arched eyebrows* classifiers, and $3e^{-5}$ for the *bags under eyes* classifier.

The **encoder** in the variational autoencoder is a series of 5 convolutional blocks. Each block shares the same structure, with *kernel* size 3, *stride* 2, *batch normalization* and *leaky-relu* activation with negative slope coefficient of 0.3. The number of filters at the output of each block is 32, 64, 128, 256 and 512. After the last block we insert a flattening layer and two dense layers each with 256 output neurons for the distributional mapping to the latent space. The **decoder** is a fully connected dense layer with 80192 output neurons (reshaped into a $4 \times 4 \times 512$ activation map) followed by 5 up-sampling blocks. Each block up-samples the input by a factor 2 and feeds it into a convolutional layer with kernel size 3 and stride 1, followed by *batch normalisation* and *leaky-relu* activation with 0.3 negative slope coefficient. The number of channels at the output of each block are 256, 128, 64, 32 and 3 respectively. We apply an additional convolutional layer with kernel size 3, stride 1, 3 output channels and *sigmoid* activation for a final reconstructed RGB image with values restricted in the $[0, 1]$ interval.

The autoencoder is trained for 100 epochs using the *Adam* optimizer, with batch size 64 and a learning rate of $5e^{-4}$ which is decreased after each epoch by a factor of 0.98. We use a *perceptual loss* function together with the Kullback-Leibler divergence regularisation term, following details on [Hou et al., 2017] (VAE-123 model).

D.4 ATTRIBUTION METHODS

We use standard implementations of attribution methods with recommended parameters in corresponding publications or public repositories. In all cases, *black+white* and

counterfactual variants of methods are implemented equivalently. For path methods requiring trapezoidal integration, we use 50 bins with grayscale images and 25 bins with high resolution images. The process to procure counterfactual fiducials is explained in Section 3.

Vanilla IG is implemented with a straight line as domain of integration.

Blur IG is specified with an integration path which decreases blurring from a masked image, using successive Gaussian filters. The maximum standard deviation is set to the minimum required to maximise the average predictive entropy across train data.

Guided IG is configured s.t. the subset of pixels traversing value in each step is the 10% with smallest partial derivatives of entropy wrt pixel values. We use 50 steps.

LIME is implemented through *quickshift* segmentation, with kernel 1, maximum distance 5 and ratio of 0.2. We use a binomial mask with deactivation probability 0.2, and *Lasso* regression to attribute importances.

SHAP proceeds through $2 * (\text{Pixel Count}) + 2^{11}$ index perturbations of varying size; masked index points are resampled from their corresponding marginal distributions. We use *Lasso* regression to attribute importances.

CLUE attributions are derived as the difference between an image and its decoded CLUE counterpart [cf. Antoran et al., 2021, Appendix F]. The cost function weighs reconstruction and uncertainty terms, and is tuned on a validation set.

Xrai is implemented with Felzenszwalb's segmentation algorithm in order to retrieve masks. We use multiple scale values of $50, 100, 150, 250, 500$ and 1200, as well as a dilation radius of 5. This is applied to normalised images at range $[-1, 1]$ and size 224×224 pixels. Resizing is undertaken with anti-aliasing. Segments are accepted for appending into attributions with a required difference of 50 pixels.